ISSN 1870-4069

Classifications and Grammars of Simple Arabic Lexical Invariants in Anticipation of an Automatic Processing of this Language "The Temporal Invariants"

Dhaou Ghoul[1], André Jaccarini[2], Amr Helmy Ibrahim[1]

[1] Sorbonne University,
STIH Laboratory,
France

[2] Aix-en-Provence University,
MMSH,
France

{dhaou.ghoul, amrhelmyibrahim, andre.jaccarini}@gmail.com

Abstract. Our study focuses on the classification and treatment of simple Arabic lexical invariants that express a temporal aspect. Our aim is to create a diagram of grammar (finite state machine) for each invariant. In this work, we limited our treatment to only 13 lexical invariants. Our hypothesis begins with the principle that the lexical invariants are located at the same structural level (formal) as the schemes in the language quotient (skeleton) of the Arabic language. They hide a great deal of information and involve syntactic expectations that makes it possible to predict the structure of the sentence. To do this: First, for each lexical invariant, we developed a sample corpus that contains the different contexts of the lexical invariant in question. Second, from this corpus, we identified the different linguistic criteria of the lexical invariant that allow us to correctly identify it. Finally, we codified this information in the form of linguistic rules in order to model it by a diagram of grammar (finite state machine).

Keywords. Corpus, classification, syntactic environment, regular expression, Arabic language, lexical invariants, identification, linguistic rules, regular grammar, diagrams of grammars, NLP.

1 Introduction

As part of Mogador[1] project [1] we are interested in tool words that we call lexical invariants. Indeed, morphosyntactic ambiguity is most often identified in the Arabic language. According to our research, we noticed that several arguments demonstrate

[1] Modélisation des grammaires arabes, des données et des outils de recherche. https://halshs.archives-ouvertes.fr/halshs-00912009/document.

the ambiguity of tokens: For example, various graphical shape due to the absence of vowels and treating the word out of context. By contrast, a mechanism for morphosyntactic disambiguation of Arabic is needed to resolve the ambiguity of unveiled words. The aim of our research is to treat the Arabic language with minimum although complex rules without referring to a lexicon. In their work Audebert *et al.* proposed grammars that define some syntactic operators in Arabic (’inna, ’anna, ’an…) [2]. Our work is based on this bibliography to improve the grammatical basis of the operators.

In general, lexical invariants are classified according to their meaning and their function in the sentence. Consequently, these invariants play an important role in the interpretation of the sentence and the coherence of a text.

Our linguistic study consists in treating the context of these lexical invariants based on a corpus that represent a sample of each lexical invariant.

This study is organized as follows: In section 2: we present the concept of “lexical invariant” by exposing the various levels of invariance. Then, we classify the simple lexical invariants according to several criteria. In section 3: we present our method of linguistic study, which includes modeling through diagrams of the grammar’s some temporal lexical invariants. In section 4: we present two examples of simple lexical invariants. Conclusions and perspectives will be presented in section 5.

2 Definition and Classification of Simple Arabic Lexical Invariants

2.1 Definition of Lexical Invariant

As part of our research project, we try to project the Arabic language on its skeleton noted L/RAC. This skeleton is a quotient language whose elements are constituted of:

– Classes of equivalence of lexemes that are the schemes (patterns, lexical paradigm, wazn).
– Singletons (classes with only one element) which we can not associate a schemes (wazn). So, these singletons represent the invariants of the Arabic language that retain their same forms in the quotient language. This is why we have called "lexical invariants".

This language is an abstract object -*formally*- constructed from the free monoid constituted by the set of concatenations of the Arabic graphemes A* (with, A: Arabic alphabet and * designating the Kleene star) which the language L_{AR} consisting of the set of licit. Arabic graphic forms (graphic words separated by two whites) is a strict subset. What we mean by LEX is a particular subset of $L_{AR} \subset A^*$. Thus, $LEX \subset L_{AR} \subset A^*$. As well as, LEX is obtained by systematic segmentation of the elements of L_{AR}. [3]. Indeed, our appellation of “lexical invariant” is inspired by the definition of André Jaccarini [4].

Table 1. Temporal lexical invariants distribution.

	#sentences	**#words**	**#lexical invariant**	**#temporal invariant**	**%temporal invariant**
Text1	142	3460	1174	173	14.3
Text2	63	1311	507	80	15.7
Text3	123	2366	752	109	14.49
Text4	11	344	108	10	9.25

Let SC be the morphism associated with the operation of projection on the scheme.

We call the lexical invariant every element x belongs to LEX that is invariant with respect to the morphism SC.

The set of lexical invariants is denoted by:

$$INL=LEX \text{ inter } \{x ; SC(x)=\{x\}\}\{x ; x \in LEX \text{ et } SC(x)=\{x\}\}. \quad (1)$$

For example, SC (فإنهم/ *fa'inahum*) = ف SC (إن) هم = هم. إن. ف . With إن ϵ LEX, So إن is a lexical invariant.

These lexical invariants play a crucial role because they are the only elements that are both in the terminal vocabulary of L and L/RAC. They are on the same level as the schemes in L/RAC (quotient language) and they escape from the rules of morphological derivation. With, L is the language which consists of the set of grammatical sentences - not necessarily endowed with meaning-.

In Arabic there are two types of lexical invariants: simple lexical invariants like "*ḥattā*" and complex lexical invariants like "*ḥīnamā*". The latter are on the form "x *mā*" (with x representing a simple lexical invariant). In this study we treat only the simple lexical invariants.

2.2 Classification of Simple Temporal Arabic Lexical Invariants

Temporal lexical invariant classification remains an indispensable step in the automatic processing of the Arabic language. Based on literature on the one hand and on our study of these invariants on the other hand, we have tried to classify these 13 simple temporal invariants. Moreover, we calculated the percentage of temporal invariant's occurrences with respect to the total number of lexical invariants in four texts selected via our working corpus, which we created in previous work *[5, 6]*. Our calculation is made with the application "*Kawâkib*" *[7]* followed by a manual check.

The following table shows the temporal lexical invariants distribution with respect to all invariants in a given text.

The aim of this classification is to show the linguistic criteria of these types of words in Arabic and their influence in the construction of a text. The list of temporal invariants that we treat in this work is as follows: *l.m.ā; 'aṯnā'a; 'iḏā; 'iḏ; ḥattā; ḥīn/ ḥīna; lam; lan; munḏu; qad; ṯumma; baʿda; sawfa.*

Our classification of simple temporal lexical invariants is based on five linguistic criteria: agglutination, based on the word that follows them, syntactic role, ambiguity and based on their *rection* in the sentence.

Table 2. Lexical invariants classification based on their agglutination.

Class	Classification criteria	Example
1	Agglutination only to a coordinating conjunction.	*l.m.ā, ʾiḏ, ḥattā*
2	Agglutination to a coordinating conjunction and / or an interrogative conjunction.	*ʾiḏā, lam, lan*
3	Agglutination to a coordinating conjunction and / or a corroborating conjunction.	*qad*
4	Agglutination to a coordinating conjunctions and / or preposition and a personal pronoun attached.	*ʾaṯnāʾa*
5	Agglutination to a coordinating conjunctions and / or preposition and a personal pronoun attached and to a relative pronoun.	*ḥīn*
6	Agglutination to a coordinating conjunction and / or a personal pronoun attached and a relative pronoun.	*baʿda*
7	Agglutination to a coordinating conjunction and / or an interrogative conjunction and a corroborating conjunction.	*Sawfa*

Table 3. Lexical invariants classification based on the word that follows them.

Class	Classification criteria	Example
1	Pro-nominal invariants	*ʾaṯnāʾa, baʿda*
2	Pre-verbal invariants	*qad, lam, lan, sawfa*
3	Mixed invariants	*l.m.ā, ḥattā, ʾiḏ, ʾiḏā, ṯumma, munḏu, ḥīna*

Lexical invariants classification based on their agglutination with the prefixes or suffixes, we classified these 13 simple lexical invariants into 7 classes as the following table 2 shows.

Lexical invariants classification based on the word that follows them. According to our study of lexical invariants, it seemed useful to classify an invariant based on the word that follows them. Therefore, based on this criterion we have classified these 13 simple lexical invariants into three classes or families: pro-nominal invariants[2], pre-verbal invariants[3], mixed invariants[4] as the following table 3 shows.

Lexical invariants classification based on their syntactic roles. In Arabic language, we can find lexical invariants whose graphic form is unique but can have several syntactic functions. Among these lexical invariants is "*ḥattā*" whose unique graphic form may contain a subordinating conjunction, coordinating conjunction, adverb or preposition. Based on their syntactic roles, these lexical invariants are divided into 9 classes as the following table 4 shows.

Lexical invariant classification based on their "*rection*" in the sentence. Some lexical invariants in Arabic change the vocalization of the word that follows them. This

[2] The lexical invariants that precede the noun or the elements equivalent to the noun.

[3] The lexical invariants that precede the verb or the elements that can receive the verb.

[4] The lexical invariants that precede either a verb or a name or another invariant.

Table 4. Lexical invariants classification based on their syntactic roles.

Class	Syntactic class	Example
1	Subordinating conjunction	*lammā, lamā, limā, ’iḏ, ḥattā, ’iḏā, ’aṯnā’a, ḥīna, munḏu*
2	Coordinating conjunction	*ḥattā, ṯumma*
3	Negation particle	*lan, lam, lammā*
4	Interrogative pronoun	*limā*
5	Preposition	*ḥīna, ’aṯnā’a, baʿd, ḥattā, munḏu*
6	Temporal adverb	*l.m.ā, ’iḏ, ’iḏā, ’aṯnā’a, ḥīna, munḏu, baʿd, sawfa, ḥattā, lan, lam, ṯumma, qad*
7	Conditional conjunction	*’iḏ, ’iḏā*
8	Surprise particle	*’iḏ, ’iḏā*
9	Future particle	*sawfa*

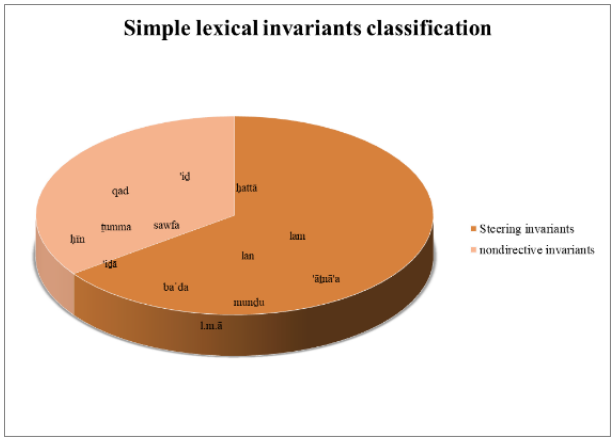

Fig. 1. Lexical invariants classification based on their “reaction” in the sentence.

linguistic criterion is useful in the identification of these lexical invariants. It is therefore important to classify these invariants according to their power of reaction. Indeed, and based on this criterion, we classified these lexical invariants into two classes as the following figure (Fig.1) shows:

Lexical invariant classification based on their ambiguity. The vowels absence in the Arabic text is one of the major problems of automatic processing of this language. This problem leads to several cases of grammatical (syntactic) and graphical or even semantic ambiguity. Moreover, the phenomenon of agglutination in Arabic can be a major source of ambiguity. Therefore, it seemed useful to classify our study invariants according to their ambiguity (graphic, syntactic and semantic) as the following table 5 shows.

It is worth noting that the different abbreviations are illustrated in the appendix. In general, the lexical invariants classification depends on their linguistic properties. Therefore, this classification is always debatable and not exhaustive.

Table 5. Lexical invariants classification based on their ambiguity.

Lexical invariant	Graphic ambiguity	Syntactic ambiguity	Semantic ambiguity
l.m.ā	*lammā, limā, lamā*	CS, PNE, ADVT	Correspondence, temporal, negation
’iḏ	*’iḏ, āḏ*	CS, COND, ADVT, PS	Condition, surprise
’iḏā	*’iḏā, āḏā, āḏan, ’iḏan*	CS, PS, CCOND, ADVT	Condition surprise
lan	Unique	ADVT, PNE	Negation
lam	Unique	ADVT, PNE	Negation
ḥattā	Unique	CS, CC, PREP, ADVT	Coordination, intensity, justification, direction…
ṯumma	*ṯumma, ṯamma*	CC, ADVT, PD, Verb	Ranking (order)
munḏu	Unique	CS, ADVT, PREP	Time report, temporal interrogation
qad	Unique	ADVT	Certainty, uncertainty
ḥīna	ḥīna, ḥīn	ADVT, CS, PREP	Temporal correspondence, a moment
’aṯnā’a	*‘aṯnā’a, aṯnā’*	CS, ADVT, PREP	Temporal synchronization
	baʿda,	PREP	Temporal, Continuity
bʿd	baʿdu,	ADVT	Anteriority
	buʿd,	Noun	Distance
	baʿida, baʿuda, baʿada	Verb	Move away
sawfa	Unique	ADVT, PF	Promise

3 Methodology of Lexical Invariants Modeling by Grammars Diagram

We will start from the idea that a grammar is a point of view among others on the language and that these points of view are related to the tasks assigned to these grammars. The grammars of the invariants are made up on them and their linguistic analysis. Information about them is largely known by Arab and Western grammarians. However, our contribution is summarized in the angle of analysis that we adopt and the examples that we select.

So, the grammars presented in our work are intended for the automated morpho-syntactic analysis of the Arabic language. The idea here is to transform the linguistic rules of each invariant that we have defined previously into a non-fixed-form formal grammar. These grammars describe the syntactic expectations of invariants in Arabic [8].

3.1 Different Levels of Our Method

Our method of Arabic lexical invariant modeling is located as part of the theory of “abstract machines”. It is based on two major phases (two levels of analysis).

The linguistic level. Our methodology of the study linguistic of each lexical invariant consists to study its environment based on a sample of representative sentences. The different steps of our method are:

- Identification of lexical invariants in the corpus.
- Treatment of the environment (left and right context) of the lexical invariant.
- Develop hypotheses on the structure of the sentence based on the principle of the "empty sentence".
- Identify the different characteristics of the lexical invariants (grammatical role, ambiguity, suppressibility or not, graphical form, distributional analysis [9, 10], local and global scope...).
- Study of the syntactic expectations of lexical invariants.
- Codify the different information into non-fixed formal grammars (rules of production or linguistic rewriting), to bring out the relations in order to make them usable by the machine.

Note that the semantic part of each invariant will be studied in parallel and can also be related to the definition of discriminating criteria.

The graphic level. This level consists in modeling the grammars in the form of finite state automata (grammars schemes) in order to make it comprehensible by a machine. The graphs represent successions of lexical invariant's syntactic expectations. This means that one passes from a state represented by a node to another state by means of a transition which is symbolized by an arc. These arcs are labeled with the terminal symbols of the grammar.

So why use the theory of automata to represent or treat the Arabic language?

Indeed, the structural (strongly grammatical) characteristics of Arabic are also algorithmic characteristics. These characteristics are easily translatable as part of the theory of automata. Thus, the theory of automata offers the possibility of a good study of Arabic in the universe of knowledge.

3.2 Limits of Our Method

In a general way, our method, purely algorithmic, differs from those of contemporary research. It uses minimal resources and is independent of lexicons. The origin of this approach is due to the representation of the Arabic language in the form of a quotient or skeleton language. The latter can be obtained by reducing all the roots into a single control root "faʿala" to represent all Arabic schemes.

Indeed, at this stage we have seen that in the language quotient exist particular words that we note by "lexical invariants". Our method consists in a surface analysis of these invariants in order to construct non-fixed formal grammars that represent the environment of each invariant. These grammars can be linked and combined. However, our method presents some limits:

- "Ad hoc" construction designed for a given task: The automata may contain unnecessary parts.

- Difficulty in marking the meaning of the lexical invariant in the automaton.
- Grammar diagrams may in some cases produce grammatical correct, but semantically incorrect sentences (invalid sentences because of the presence of a succession of epsilons).

4 Examples of Simple Temporal Lexical Invariants

In this section we will present two examples of simple lexical invariants among the 13 invariants that we have analyzed in this work. First of all, we will extract the maximum of information in the form of linguistic rules. Then, we will model these rules in regular grammars acceptable by finite state automata.

4.1 Lexical Invariant "*ḥīn /ḥīna*"

Outside of its context and in the absence of vowels, the word "*ḥīn*" is ambiguous grammatically. Indeed, this word may be a temporal subordinating conjunction or a noun. The disambiguation of this word depends on its position in the sentence as well as on the linguistic study of its environment. To assign the correct grammatical function to "*ḥīn*", we must treat its context based on a sample corpus. This sample corpus represents an extract of 60 sentences from our working corpus [5, 6].

Independently of its grammatical function, the removal of "ḥīn" causes the disruption of the structure as well as the meaning of the sentence.

Temporal subordinating conjunction. In this case, the lexical invariant "*ḥīn*" expresses a relation of time between two events and noted "*ḥīna*". This is a link between two parallel events as shown in the following example (1). In addition, it is very fertile lexical invariant. That is, it can agglutinate at n + 1 with suffixes like "*mā*," "*iḏan*", "*ḏāk*" to establish a new lexical invariant.

أصابته الدهشة حين زار مدينة القاهرة » *'asābathu aldahšat hīna zāra madīnat alqāhirat*.» (He was astonished when he visited the city of Cairo). (1)

أصابته الدهشة عندما زار مدينة القاهرة » *'asābathu aldahšat ʿindamā zāra madīnat alqāhirat*. » (He was astonished when he visited the city of Cairo). (1')

From the examples (1) and (1') above, we note that we can replace "*ḥīna*" by "*ʿindamā*" without any change in the sentence either in the sense or in the structure. However, this permutation is only valid in the case where "*ḥīna*" has the same syntactic expectations (a verb phrase) as "*ʿindamā* "(sentence 1'). This permutation is strictly forbidden in the case where "*ḥīna*" is followed by a nominal group as shown in the example (2):

في حين أن الكثير من الشباب ملتزمون بجلسات المقاهي » *fī ḥīna 'anna alkaṯīr min alšabāb multazimūn bigalasāt almaqāhī*.» (However, many young people were keen on going to coffees). (2)

If we look at the example (2), the sequence "*fī ḥīna*" is equivalent to "*baynamā*" (2') which marks a temporal opposition between two non-identical events. So, the lexical invariant "*ḥīna*" does not keep the same meaning when it appears in a sequence of invariants like "*fī ḥīna*". It no longer expresses temporality, but it expresses an opposition.

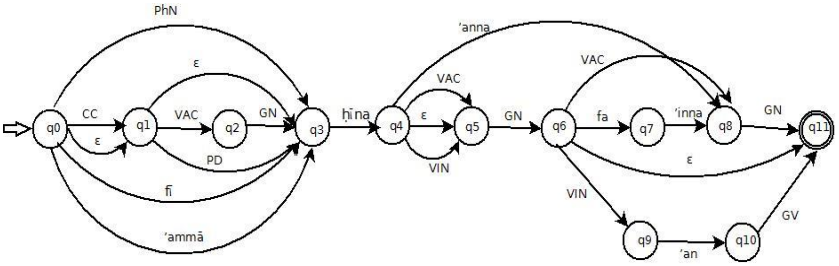

Fig. 2. Grammar diagram of "*ḥīna*" temporal subordinating conjunction.

بينما أن الكثير من الشباب ملتزمون بجلسات المقاهي « *baynamā 'anna alkaṯīr min alšabāb multazimūn bigalasāt almaqāhī*.» (By contrast, many young people regularly attended cafés). (2')

As for syntactic expectations, the lexical invariant "ḥīna" in this case, waits in n +1 either a verb phrase (the verb can be either in the present tense or in the past tense), or a nominal group. However, it can be preceded in n-1 by a coordinating conjunction, the preposition "*fī*", a demonstrative pronoun, corroborating conjunction "*ammā*" or the preposition "*li*". Note that in general if this invariant preceded in n-1 by the invariant "*fī*", it will be followed in n + 1 by "'*anna*".

The Formal grammar that accepts phrases in which "*ḥīna*" plays the role of a temporal subordinating conjunction is presented in the form of a finite automaton comprising 12 states as the following figure (Fig.2) shows. G (*ḥīna*, CST)= {N, ∑, P, q0, q11}, with N= {q0, q1, q2, q3, q4, q5, q6, q7, q8, q9,q10, q11} is the not terminal vocabulary, ∑= {VAC, VIN, GN, GV, CC, PD, PhN, *ḥīna*, *'inna*, *'anna*, *'an*, *'ammā*, *fī*, *fa*, ε} is the input vocabulary, P is the set of production rules and q0ϵN is an initial state, q11ϵN is the final state.

G (*ḥīna*, CST) =

1. q0→ CC q1| PhN q3|fī q3|'ammā q3|ε q3,
2. q1→ VAC q2|PD q3|ε q3,
3. q2→ GN q3,
4. q3→ ḥīna q4,
5. q4→ VAC q5| VIN q5|'anna q8|ε 5,
6. q5→ GN q6,
7. q6→ VAC q8|VIN q9|fa q7|ε q11,
8. q7→ 'inna q8,
9. q8→ GN q11,
10. q9→ 'an q10,
11. q10→ GV q11.

Noun (ḥīn). In general, the lexical invariant "*ḥīn*" occupies the function of a name in the case where it intervenes either in the middle or at the end of the sentence as shown in the following examples (3) and (4):

ويراك في كل حين وكل لحظة « *wa yarāka fī kulli ḥīn wa kulli laḥẓat*.» (He always saw you at all times). (3)

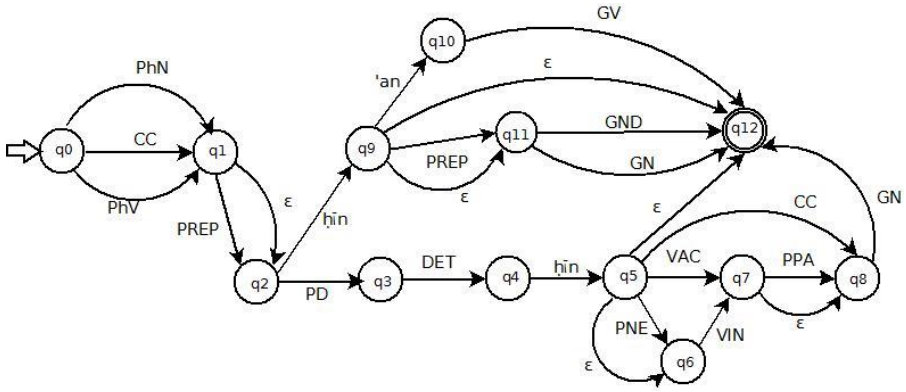

Fig. 3. Grammar diagram of "*ḥīn*" noun.

ولتعلمن نبأه بعد حين « *wa litaʿlamanna nabaʾahu baʿda ḥīn*. » (You will soon hear about it). (4)

According to the two sentences above (3 and 4), we note that "*ḥīn*" expresses the notion of a short or long moment. That is why we can be replaced it by the word "*waqt*" (moment). So in this case, "*ḥīn*" has a syntactic and semantic equivalence with the word "*waqt*" as shown in the examples (3') and (4').

ويراك في كل وقت وكل لحظة « *wa yarāka fī kulli waqt wa kulli laḥẓat*.» (He always saw you at all times). (3')

ولتعلمن نبأه بعد وقت « *wa litaʿlamanna nabaʾahu baʿda waqt*. » (You will soon hear about it). (4')

Note that in this case "*ḥīn*" can concatenate with the article "*Al*" to transform it from a noun to a determined noun as shown in the following sentence (5) or with the preposition "*li*" (example 6).

فحتى ذلك الحين أترككم في رعاية الله « *faḥattā ḏalika Alḥīn ʾatrukukum fī riʿāyat allah*. »(And in the meantime, may God protect you). (5)

ينتظرون لحين ان يأتي دورهم في الحكاية « *yantaẓirūn liḥīn ʾan yaʾtī dawruhum fī Alḥikāyat*. » (They were waiting for a moment to come to their role in history). (6)

Similarly, in this sentence (5) if the invariant "ḥīn" is replaced by the word "waqt", the structure and meaning of the sentence do not move (Example 5'). Consequently, we can permute the sequence "liḥīn" with the lexical invariant "ilā" or "ḥattā" without any change in the source sentence (example 6').

فحتى ذلك الوقت أترككم في رعاية الله « *faḥattā ḏalika Alwaqt ʾatrukukum fī riʿāyat allah*. » (And until that time God bless you). (5')

ينتظرون إلى ان يأتي دورهم في الحكاية « *yantaẓirūn ʿilā ʾan yaʾtī dawruhum fī Alḥikāyat*. » (They will wait for the time that will come their role in the history). (6')

In this case, the syntactic expectations of "*ḥīn*" are less complex than those of a subordinating conjunction. Indeed, it expects in n + 1 either the empty set or a nominal group. This nominal group can be in the form "*min* GND" (example 7).

هل اتى على الانسان حين من الدهر « *hal ʾatā ʿalā Alʾinsān ḥīn min Aldahr*. » (Did the man go through a time lapse). (7)

The Formal grammar that accepts phrases in which "*ḥīna*" plays the role of a temporal subordinating conjunction is presented in the form of a finite automaton

Table 6. Linguistic Properties of "*ḥīn*".

ḥīn /ḥīna		
Graphic form / grammatical class	ḥīna /CST	ḥīn / Noun
Meaning	Temporal: *fī alwaqt allaḏī*	Expresses the notion of a moment
Removal	No	No
Position in the sentence	Beginning or middle	Middle or end
Change of position	Yes	No
Agglutination in n-1	*wa, fa*	*li, Al*
Agglutination in n+1	*mā, ḏāk, ’iḏan*	No
Form of sentence	*ḥīn* P Q ou Q *ḥīn* P	P *ḥīn* ou P *ḥīn* Q
Switching with other invariants	*ʿindamā* *fi ḥīna ’anna =baynamā* *liḥīn= ḥattā*	*waqt*[5] *liḥīn= ʿilā*
Syntactic expectations at n + 1	GV or GN	GN, min GND or Ø
Governance	Governs two kernels: verb + subject and / or theme + predicate.	No
Discontinuous	No	No
Traduction	When	Moment

comprising 13 states as the following figure (Fig.3) shows. G (*ḥīn*, Noun)= {N, ∑, P, q0, q12}, with, N= {q0, q1, q2, q3, q4, q5, q6, q7, q8, q9, q10, q11, q12} is the not terminal vocabulary, ∑= {VAC, VIN, GN, GND, GV, CC, PD, DET, PREP, PNE, PPA, PhN, PhV, *ḥīn*, *’an*, ε} is the input vocabulary, P is the set of production rules and q0ϵN is an initial state, q12ϵN is the final state.

G (*ḥīn*, Noun) =

1) q0→PhN q1| PhV q1| CC q1,
2) q1→ PREP q2|ε q2,
3) q2→PD q3|ḥīn q9,
4) q3→DET q4,
5) q4→ ḥīn q5,
6) q5→ VAC q7|PNE q6|CC q8|ε q6|ε q12,
7) q6→ VIN q7,
8) q7→ PPA q8|ε q8,
9) q8→GN q12,
10) q9→ PREP q11| ’an q10|ε q11|ε q12,
11) q11→ GV q12.

The different linguistic properties of "*ḥīn*/ *ḥīna*" are summarized in the table below:

[55] In this case, the word "waqt" does not designate a lexical invariant but rather a noun.

Table 7. Linguistic Properties of "*ʾaṯnāʾa*".

ʾaṯnāʾa	
Graphic ambiguity	Unique :أثناء «*ʾaṯnāʾa*», اثناء «*aṯnāʾa*»
Grammatical ambiguity	Unique: Preposition
Meaning	Temporal: *ḫilāla*
Removal	No
Position in the sentence	Beginning or middle
Change of position	Yes
Agglutination in n-1	*wa, fa, ka*
Agglutination in n+1	Attached personal pronoun: *h, hā…*
Form of sentence	*ʾaṯnāʾa* P Q ou Q *ʾaṯnāʾa* P
Switching with other invariants	Yes: *ḫilāla*
Syntactic expectations at n + 1	GND
Governance	No
Discontinuous	No
Traduction	During

4.2 The Lexical Invariant "*ʾaṯnāʾa*"

According to Arabic literature, the word "*ʾaṯnāʾa*" is a preposition that connects two or more events in the sense of "*ḫilāla*". These events are parallel and simultaneous with respect to time. If we take into account how to transcribe the "Hamza", the lexical invariant " *ʾaṯnāʾa* " is not graphically ambiguous.

The aim of our study is to have more or less precise idea on this lexical invariant. To do this, we first selected some samples of sentences that contain " *ʾaṯnāʾa* " from our corpus [5, 6]. Our study sample of " *ʾaṯnāʾa* " contains 50 sentences. We used this sample at first to identify the notable features. Among these pertinent remarks, we found that " *ʾaṯnāʾa* " was always followed by a determined nominal group.

Indeed, to better understand the functioning of this invariant in the Arabic language, we have treated some selected examples from our study sample.

Consider the following examples:

وأثناء دراستي في بريطانيا قدمنا عملا كنت أعمل فيه مساعد مخرج » *wa ʾaṯnāʾa dirāsatī fī brīṭānyā qaddamnā ʿamalāa kuntu ʾaʿmal fīh musāʿid muḫrig* ». (During my studies in the United Kingdom, we presented a work in which I was an assistant director) (7).

يأخذ الرواية معه أثناء سفره إلي الخارج » *yaʾḫuḏu alriwāyat maʿahu ʾaṯnāʾa safarihi īllā alḫārig*» (He took the novel with him during his travels abroad). (8)

From examples above, we note that " *ʾaṯnāʾa* " can intervene either at the beginning or in the middle of the sentence and never at the end of the sentence except in the case where it is agglutinated in n + 1 with a Personal pronoun attached as shown in the following sentence:

فليحرص المسلم النبيه النزيه على محاسبة نفسه المحاسبة الشاملة قبل العمل وأثناءه» *faliyaḥuṣ almuslim alnabīh alnazīh ʿalā muḥāsabat nafsihi almuḥāsabat alšāmilat qabla alʿamal wa ʾaṯnāʾahu* » (The true Muslim must be vigilant in asking for complete accounts before and during the work). (9)

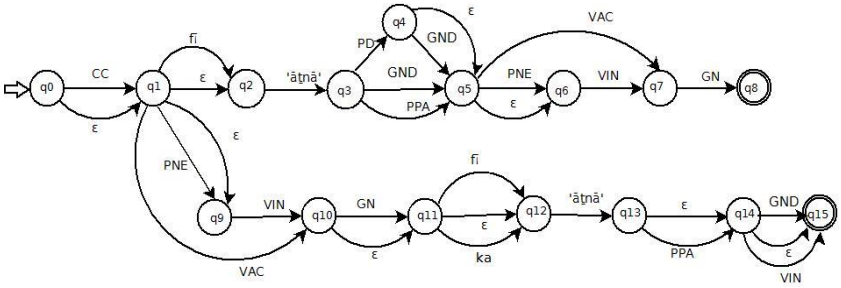

Fig. 4. Grammar diagram of "*'aṯnā'a*" *subordinating conjunction.*

As we noted above, this invariant takes the meaning of "*ḫilāla*". That is, we can switch "*'aṯnā'a"* by *"ḫilāla"* and keeping the same meaning and structure of the sentence as shown in the following example:

يأخذ الرواية معه **أثناء** سفره إلي الخارج ↔ يأخذ الرواية معه **خلال** سفره إلي الخارج (He took the novel with him during his travels abroad).

The table below summarizes these different linguistics properties:

The Formal grammar that accepts phrases in which "*'aṯnā'a* " plays the role of a temporal subordinating conjunction is presented in the form of a finite automaton comprising 16 states as the following figure (Fig.4) shows. G (*'aṯnā'a*, CS)= {N, ∑, P, q0, q15}, with N= {q0, q1, q2, q3, q4, q5, q6, q7, q8, q9, q10, q11, q12, q13, q14, q15} is the not terminal vocabulary, ∑= {VAC, VIN, GN, CC,PPA, PNE, PD, fī, ka, 'aṯnā'a, ε} is the input vocabulary, P is the set of production rules and q0ϵN is an initial state, q15ϵN is the final state.

G (*'aṯnā'a*, CS) =

1) q0→ CC q1| ε q1,
2) q1→ fī q2 | VAC q10| PNE q9|ε q2|ε q9,
3) q2→ 'aṯnā' q3,
4) q3→ PD q4|GND q5| PPA q5,
5) q4→ GND q5|ε q5,
6) q5→ PNE q6| VAC q7|ε q6,
7) q6→ VIN q7,
8) q7→ GN q8,
9) q9→ VIN q10,
10) q10→ GN q11|ε q11,
11) q11→ fī q12| ka q12|ε q12,
12) q12→ 'aṯnā' q13,
13) q13→ PPA q14|ε q14,
14) q14→ GND q15| VIN q15| ε q15.

5 Conclusions and Perspectives

In this work, we tried to study and classify 13 simple Arabic lexical invariants. In our study, we have focused on the notions of "permutation" or "commutation" that based the approach of structural linguistics called "distributional". To do this, first for each lexical invariant, we developed a sample corpus that contains different contexts of the

lexical invariant in question. Second, from this corpus, we identified the different linguistic criteria of the lexical invariant that allow us to correctly identify it.
Finally, we codified this information in the form of linguistic rules in order to model it by diagram of grammar (finite state machine). Indeed, as perspectives for this work, we will evaluate our temporal invariant grammars on a big data.

6 Appendix

VIN: Verb present tense, VAC: verb past tense, VI: imperative verb, N: Noun, PREP: preposition, CC: coordinating conjunction, CS: subordinating conjunction, ADVT: adverb of time, PNE: negation particle, CCOND: conditional conjunction, CST: temporal subordinating conjunction, PS: Particle of surprise, PD: demonstrative pronoun, PF: Future particle, PPA: Pronoun personal attached. GN: Nominal group, GND: Determined nominal group, GV: verbal group, PhN: Nominal sentence, PhV: verbal sentence, DET: determiner.

References

1. Jaccarini, A., Gaubert, C.: Le programme Mogador en linguistique formelle arabe et ses applications dans le domaine de la recherche et du filtrage sémantique (2012)
2. Audebert, C., Jaccarini, A.: À la recherche du khabar, outils en vue de l'établissement d'un programme d'enseignement assisté par ordinateur. Institut français d'archéologie orientale du Caire, pp. 217–256 (1986)
3. Ghoul, D.: Classifications et grammaires des invariants lexicaux arabes en prévision d'un traitement informatique de cette langue les invariants temporels. Doctorat (2016)
4. Jaccarini, A.: Grammaires modulaires del'arabe. Mise en œuvre informatique et stratégies. Thèse de doctorat, Sorbonne (1997)
5. Ghoul, D.: Construction d'un corpus arabe à partir du Web dans le but d'identifier les mots-outils ou tokens. JADT: Journées internationales d'Analyse statistique des Données Textuelles, INALCO, pp. 271–276 (2014)
6. Ghoul, D., Ibrahim, A. H.: Web arabic corpus : construction d'un large corpus arabe annoté morpho syntaxiquement à partir du Web. CECTAL'15, pp. 12–16 (2015)
7. Gaubert, C.: Kawâkib, une application pour le traitement automatique de textes arabes. no. 44, pp. 66–81 (2010)
8. Ghoul, D, Ibrahim, A. H., Audebert, C.: Rules-based grammatical and semantic disambiguation of the token "hatta". In: 5th International Conference on Information Communication Technology and Accessibility (ICTA), pp. 1–6 (2015)
9. Dubois, J., Dubois-Charlier, F.: Principes et méthode de l'analyse distributionnelle, Langages, vol. 5, no. 20, pp. 3–13 (1970)
10. Dubois, J.: Grammaire distributionnelle, vol. 1, no. 1, pp. 41–48 (1969)